Lawrence Berkeley National Laboratory
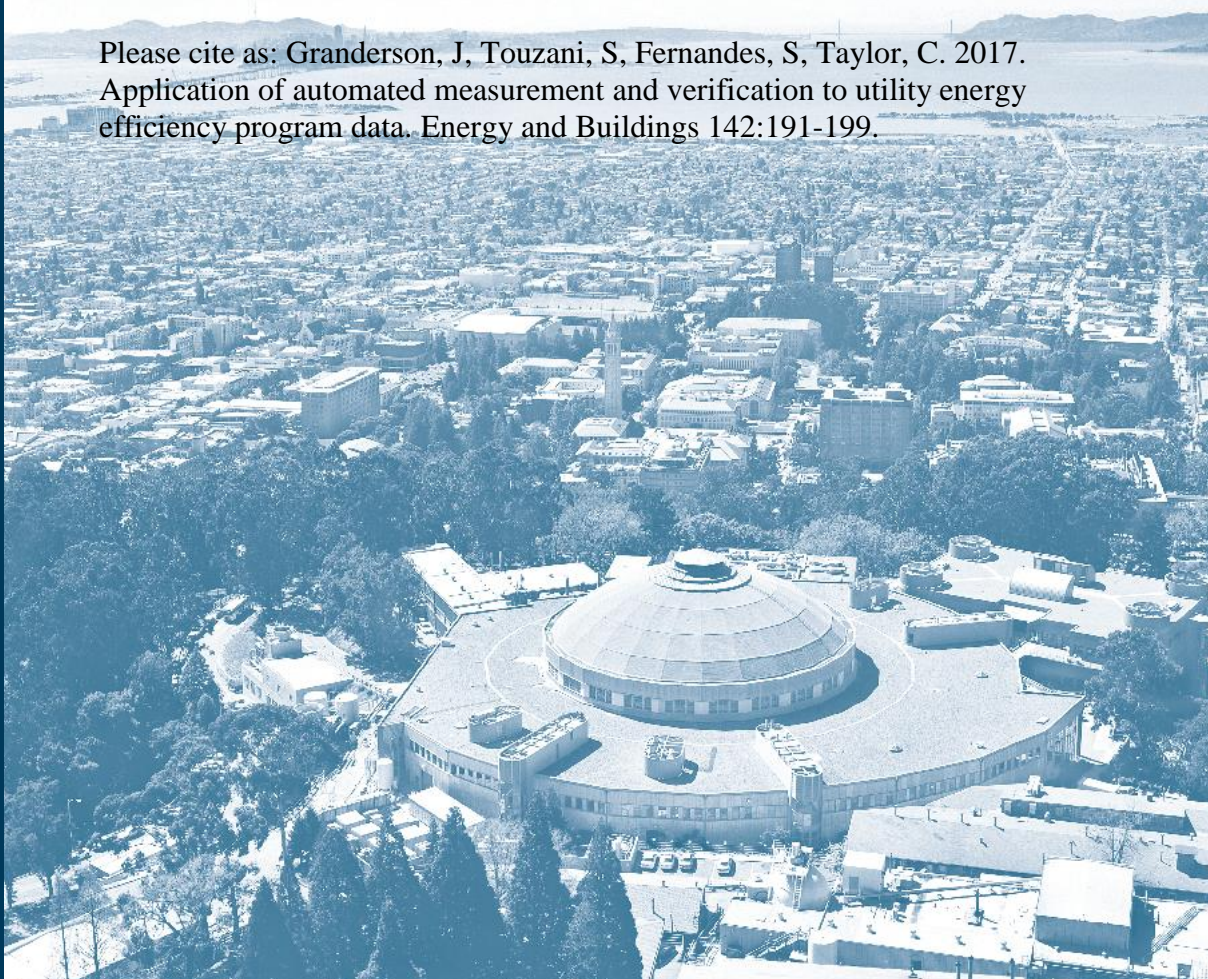
Application of Automated Measurement and Verification to Utility Energy Efficiency Program Data

Jessica Granderson, Samir Touzani,
Samuel Fernandes, Cody Taylor

Disclaimer:

Application of Automated Measurement and Verification to Utility Energy Efficiency Program Data

Jessica Granderson[*,a], Samir Touzani[a], Samuel Fernandes[a], Cody Taylor[b]

[a] Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley CA, 94720, USA
b US Department of Energy, Office of Energy Policy and Systems Analysis, 1000 Independence Ave. SW, Washington, DC, 20585, USA

ABSTRACT

Trustworthy savings calculations are critical to convincing regulators of both the cost-effectiveness of energy efficiency program investments and their ability to defer supply-side capital investments. Today's methods for measurement and verification (M&V) of energy savings constitute a significant portion of the total costs of energy efficiency programs. They also require time-consuming data acquisition. A spectrum of savings calculation approaches is used, with some relying more heavily on measured data and others relying more heavily on estimated, modeled, or stipulated data.

The increasing availability of "smart" meters and devices that report near-real time data, combined with new analytical approaches to quantify savings, offers the potential to conduct M&V more quickly and at lower cost, with comparable or improved accuracy. Commercial energy management and information systems (EMIS) technologies are beginning to offer these 'M&V 2.0' capabilities, and program administrators want to understand how they might assist programs in quickly and accurately measuring energy savings. This paper presents the results of recent testing of the ability to use automation to streamline the M&V process.

In this paper, we apply an automated whole-building M&V tool to historic data sets from energy efficiency programs to begin to explore the accuracy, cost, and time trade-offs between more traditional M&V, and these emerging streamlined methods that use high-resolution energy data and automated computational intelligence. For the data sets studied we evaluate the fraction of buildings that are well suited to automated baseline characterization, the uncertainty in gross savings that is due to M&V 2.0 tools' model error, and indications of labor time savings, and how the automated savings results compare to prior, traditionally determined savings results. The results show that 70% of the buildings were well suited to the automated approach. In a majority of the cases (80%) savings and uncertainties for each individual building were quantified to levels above the criteria in ASHRAE Guideline 14. In addition the findings suggest that M&V 2.0 methods may also offer time-savings relative to traditional approaches. Finally we discuss the implications of these findings relative to the potential evolution of M&V, and pilots currently being launched to test how M&V automation can be integrated into ratepayer-funded programs and professional implementation and evaluation practice.

Keywords: M&V 2.0, savings estimation, measurement and verification, accuracy, utility programs, energy management and information systems, automation

---

[*] Corresponding Author: 1 Cyclotron Rd., MS 90-3111, Berkeley CA, 94720; Telephone (510) 486.6792; Email address JGranderson@lbl.gov

Additional author email addresses: stouzani@lbl.gov, sgfernandes@lbl.gov, Cody.Taylor@ee.doe.gov

1. Introduction

Energy Management and Information Systems (EMIS) span a family of technologies including energy information systems (EIS), building automation systems, fault detection and diagnostics, and monthly energy analysis tools. These tools have enabled whole-building energy savings of up to 10-20% with rapid paybacks often of under three years [1,2] through multiple strategies such as: identification of operational efficiency improvement opportunities, fault and energy anomaly detection, and inducement of behavioral change among occupants and operations personnel.

Although historically targeted for use by commercial building owners, managers, and operators, these energy data analytics technologies have begun to provide functionality to support the needs of utility efficiency program administrators by segmenting, targeting, and engaging customers for enhanced program delivery. In addition to *enabling* operational savings, EMIS and related analytics tools have begun to automate the quantification of whole-building energy savings, relative to a baseline period, using empirical baseline models that relate energy consumption to key influencing parameters, such as ambient weather conditions and building operation schedule [3,4,5,6,7]. The term "M&V 2.0" is being used to describe the use of large data sets (either high frequency interval data, or large volumes across many buildings) combined with computational automation to streamline and scale the M&V process. M&V 2.0 is currently the topic of much discussion in the energy efficiency program industry, spanning emerging technology, implementation, administration and evaluation [8,9,10]. These increasingly automated approaches to M&V offer several potential benefits. The opportunity to deliver continuous savings feedback can enhance customer experience and enable identification of under performing projects. Similarly, the ability to leverage automation to streamline the savings estimation process can support cost effective scaling of efficiency programs. In addition, the industry has expressed interest in how to maximize the value of investments in advanced metering infrastructure, and how meter-based savings estimation can inform the development of deemed savings estimates.

These technologies vary from semi-automated to fully automated, and in their precise implementation. However, they often preclude the need to "hands-on" directly manipulate and import meter data for analysis, to tailor-make a model unique to each building, to run models to generate predictions, and to compute savings. Although analytics-based M&V holds great promise, several questions remain to be answered before energy managers and utility programs can confidently adopt these emerging techniques. Prior work has addressed how the baseline models in automated M&V tools can be objectively evaluated to determine their *overall* predictive accuracy, and how public and proprietary models can be tested and compared to one another. However, the industry has not yet determined how these automated approaches can be practically incorporated into practitioner work streams, nor thoroughly assessed the accuracy, cost and time trade-offs with respect to more traditional approaches. In this paper we begin to address some of these outstanding issues, focusing on the following:

For what fraction of buildings can baseline energy use be characterized with automated tools that rely on commonly available meter and weather data?

When applying automated tools, what is the range of uncertainty and confidence at which savings can be estimated? How does uncertainty improve as buildings are aggregated to determine total savings for a cohort?

How do the results from automated savings estimation compare with traditional approaches to M&V that are based on custom engineering calculations?

What are the potential time savings that can be gained by leveraging a higher degree of automation in the M&V process.

In the following, we detail how these questions were explored, using historic energy efficiency program data, in partnership with North American utility administrators, implementers, and regulators. We conclude with a discussion of the implications of this work and research pilots currently being launched to test how M&V automation can be integrated into ratepayer-funded programs and professional implementation and evaluation practice.


2. Review of Prior Work

2.1 General Baseline Model Performance Assessment

Baseline energy use models characterize building load or consumption according to key explanatory variables such as time-of-day and weather. These baseline models are used for a variety of purposes in EMIS, including near real-time energy anomaly detection, and near future load forecasting, as well as quantification of energy or demand savings. Baseline model accuracy is critical to the accuracy of energy savings that are calculated according to the IPMVP. For both whole-building and measure isolation approaches (IPMVP Options B and C) the baseline model is created during the "pre-measure" period, before an efficiency improvement is made. The baseline model is then projected into the "post-measure" period, and energy savings are calculated based on the difference between the projected baseline and the actual metered use during the post-measure period [11].

Prior work established a 4-step statistical procedure that can be used to evaluate the predictive accuracy, of a given baseline model [12,3]. The test dataset comprised interval meter data and independent variable data, such as outside air temperature, for dozens to hundreds of buildings. These buildings were "untreated" in terms of efficiency interventions. That is, they are not known to have implemented major efficiency measures. The data for each building is divided into hypothetical baseline (i.e., model training) periods and hypothetical post-measure (i.e. model prediction) periods. Meter data from the prediction period is "hidden" from the model. The trained, or fitted model is used to forecast the load throughout the prediction period, and predictions are then compared to the actual meter data that had been hidden. Figure 1 shows an example of actual, and model-predicted data for a 12-month training period and a 12-month prediction period. Performance metrics that quantify the difference between the model prediction and the actual load are calculated and used to characterize accuracy.
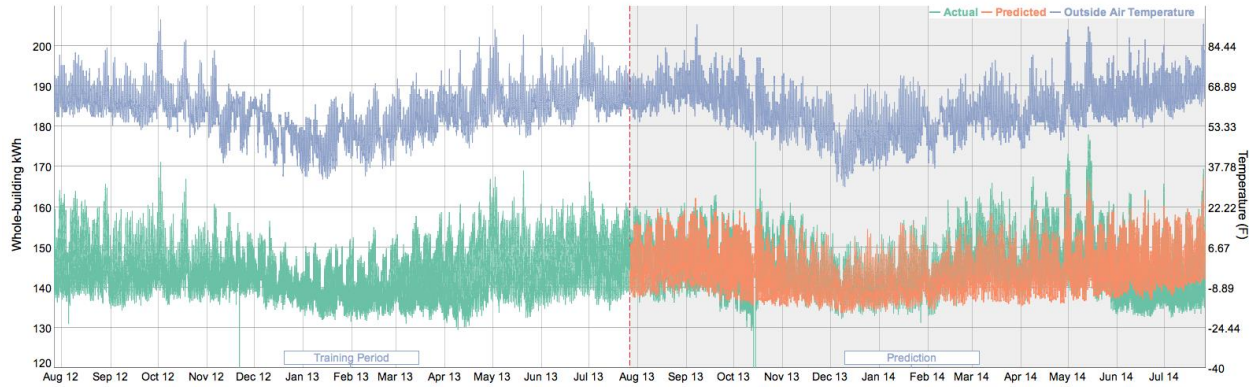
Figure 1. Actual and model-predicted energy data, overlaid with outside air temperature, for a 12-month training period and 12-month prediction period.

This testing procedure assesses model performance in general, 'on average' across populations of many buildings. Once it has been determined that a new model or tool of interest performs well *on a population basis*, the practitioner can have reasonable confidence that the model is viable for use in M&V. The tool can then be applied to *specific* buildings that have undergone, or will undergo an efficiency intervention. The principles of ASHRAE Guideline 14 [13] may be followed to determine whether the model fit is sufficient for each building in the program, and to quantify the uncertainty in the savings that are estimated. Savings and uncertainties for each building can be aggregated for a portfolio level analysis. This approach is described further in the Methodology, Section 3.

2.2 Assessment of Open and Proprietary Models

The testing procedure described in Section 2.1 was applied to evaluate a set of ten open and proprietary M&V2.0 interval baseline models [4]. The test data set comprised over five hundred buildings, primarily from California, Washington, DC, and Seattle. Overall, the models that were tested were able to predict whole-building energy use with a high degree of accuracy for a large portion of the buildings in the test dataset. For the standard whole-building case of twelve months training followed by twelve months of prediction the models were able to *predict* (as opposed to *fit*) a year of energy use with normalized mean bias error (NMBE) ranging from -1% to 4%, for one quarter of the buildings in the data set and approximately -1% to -5% for another quarter. When the models' relative performance was compared using both NMBE and CV(RMSE) for predicted energy use, the observed differences were generally quite small, with no clear 'winner' in the group.

3. Methodology

The prior work described in Section 2 established that automated baseline models can be tested for their predictive accuracy, and that on a population basis they are compellingly accurate for M&V applications. Given these encouraging findings, the authors initiated the current study to begin evaluating the performance of automated M&V tools when applied to data from *real-world utility programs*. With three North American utility and implementer partners, from the west, the northwest, and the northeast, a data set from previously completed efficiency projects

was compiled for analysis. In each case the partner was able to provide whole-building 15-minute interval electricity meter data for buildings that participated in commercial programs. However as summarized in Table 1, each program data set consisted of different supplementary information, and was provided in different formats.

Table 1. Summary of data sets contributed by utility partners, and used in the analysis

Data Set ID	Measure Types	Previously Calculated Gross Savings	Estimated Time to Calculate Gross Savings	Number of Buildings for Which Data Was Provided	Number of Buildings With At Least 9 mo. Baseline and Post-measure Data
Data Set 1	Retro-commissioning with some retrofit	Not available	Provided by partner	229	51
Data Set 2	Unknown	Provided by partner	Provided by partner	14	10
Data Set 3	Custom	Provided by partner	NA	54	23

Since the data came from previously evaluated programs that used alternate M&V approaches (such as custom engineering calculations based on shorter-term equipment- or system-level measurements), the number of buildings for which data was *provided* was much larger than the number of buildings that could be analyzed with an whole-building approach. Since the projects were not originally conducted with the intent of using a whole-building M&V approach, the historic data record did not always include the industry-standard twelve months of pre- and post-measure data. For this study buildings with less than nine months of pre- and post-data were excluded from the analysis, resulting in an overall data set that comprised eighty-four buildings. The 9-month threshold was set based on prior work that showed that 12 months of energy consumption can, in general, be predicted with small error, based on 9 months of hourly training data [3,4]. Common practice in the industry for whole-building M&V is to use 12 months of data for both the pre- and post- periods, however, this may be an artifact of historically having access exclusively to monthly whole-building data.

The automated M&V tool that was used in this work incorporates the *time of week and temperature* baseline model that was described and tested alongside nine other models in Granderson [4], demonstrating reliable predictive accuracy. The time of week and temperature model was selected for this study for two primary reasons: it is an open-source automated baseline model to which the authors have ready access, and it was shown to have predictive accuracy on par with the other models tested in prior work. In this model the predicted load is a sum of two terms: (1) a "time of week effect" that allows each time of the week to have a different predicted load from the others, and (2) a piecewise-continuous effect of temperature. The temperature effect is estimated separately for periods of the day with high and low load, to capture different temperature slopes for occupied and unoccupied building modes. The M&V tool is programmed to automatically compute avoided energy use, according to the principles of the IPMVP Option C [11]. That is, a baseline model is fit to the meter data from the pre-measure baseline period; the model is forward projected to generate energy use predictions for the measure post period; and the energy savings is computed as the difference between the model-

predicted energy use and the metered energy use. The tool is also programmed to compute savings uncertainty due to baseline model error, using the fractional savings uncertainty (FSU) metric defined in ASHRAE Guideline 14 [13]. This metric is defined as the ratio of the savings uncertainty to the savings. For example, if the estimated savings is equal to 10,000 kWh and the corresponding savings uncertainty is equal to 1,000 kWh at 95% confidence level, then the fractional savings uncertainty would be equal to 10%. Since auto-correlated model errors arise when high resolution metered data is used (e.g., 15-min, hourly or daily), ASHRAE's guideline introduced a corrected version of the FSU equation. This correction is done in a deterministic manner by reducing the number of the independent observations depending on the lag of autocorrelation. Thus the number of independent observations, n', of n observations having a lag 1 autocorrelation equal to ρ is defined as:

$$n' = n \cdot \frac{1 - \rho}{1 + \rho}$$

Thus the FSU formulation is expressed as:

$$\frac{\Delta E_{save}}{E_{save}} = \frac{1.26 \cdot t_\alpha}{m \cdot \bar{E}_{pre} \cdot F} \sqrt{MSE \cdot \frac{n}{n'} \cdot (1 + \frac{2}{n'}) \cdot m}$$

where:

- E_{save} is the estimated energy savings in the post-retrofit period
- ΔE_{save} is the uncertainty in the savings
- n is the number of periods (data points) in the pre-retrofit period
- m is the number of periods in the post-retrofit period
- $\bar{E}_{pre}$ is the mean of the actual energy consumption in the pre-retrofit
- F is the fractional savings defined as

$$F = \frac{\hat{E}_{post} - E_{post}}{\hat{E}_{post}} = \frac{E_{save}}{\hat{E}_{post}}$$

  with $\hat{E}_{post}$ is the estimated energy consumption in the post-retrofit period, E_{post} is the actual energy consumption in the post-retrofit period
- t_α is the t-statistic value with α confidence level
- MSE is the mean squared error of the baseline regression model defined as

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(E_{pre} - \hat{E}_{pre}\right)^2$$

  with E_{pre} is the actual energy consumption is the pre-retrofit period and $\hat{E}_{pre}$ is the estimated energy consumption in the pre-retrofit period

This FSU formulation expresses the uncertainty in the calculated savings as a function of the number of data points in the baseline and post-measure period, the model MSE from the baseline period, the desired confidence level, and the magnitude of savings that is calculated. In addition, for the FSU equation it is straightforward to define the uncertainty in the savings as:

$$\Delta E_{save} = 1.26 \cdot t_\alpha \cdot \frac{\hat{E}_{post}}{m \cdot \bar{E}_{pre}} \sqrt{MSE \cdot \frac{n}{n'} \cdot (1 + \frac{2}{n'}) \cdot m}$$

The savings and the corresponding uncertainty of individual buildings can be aggregated to quantify the results at a portfolio level. By supposing that the results for each building are statistically independent, the FSU for a portfolio is defined as:

$$\frac{\Delta E_{save}^{portfolio}}{E_{save}^{portfolio}} = \frac{\sqrt{\sum_{i=1}^{N}(\Delta E_{save}^i)^2}}{\sum_{i=1}^{N} E_{save}^i}$$

In this equation, E_{save}^i is the estimated energy savings in the post-retrofit period for building i and ΔE_{save}^i is the corresponding uncertainty in the savings.

While *many* buildings can be well modeled with automated modeling approaches that are based commonly available data such as time of day, day of week, and outside air temperature, some buildings require a more customized approach with additional variables, or more customized modeling solutions. It is therefore necessary to establish a process that enables a practitioner to leverage the streamlining benefits of automation, while still maintaining a quality result. In this work, the following process was applied, drawing heavily from, and extending the best practices defined in ASHRAE Guideline 14:

1. Automatically fit the model to the data from the measure pre-period, i.e. the baseline period.
2. Compute goodness of fit metrics R^2, CV(RMSE) and NMBE (for definitions of these metrics, refer to the Appendix) from the pre-period meter data, and the fit model[†].
3. Set aside the buildings for which the NMBE was greater than 0.5%, the R^2 was below 0.6 and/or CV(RMSE) surpassed 25%; these will require inspection of the data, and engineering expertise to determine whether a better fit can be obtained through adjustments or tailoring.
4. For the remaining buildings automatically compute savings according to IPMVP Option C.
   a. For each building, estimate the uncertainty in the calculated savings according the ASHRAE Guideline 14.

[†] R^2 is the coefficient of determination, CV(RMSE) is the coefficient of variation, and NMBE is the normalized mean bias error. These metrics are used to characterize different aspects of model error. Formulas to compute these metrics can be found in common statistical references, and are provided in the Appendix; CV(RMSE) and NMBE are described in ASHRAE Guideline 14 [13].

5. Aggregate the savings and the uncertainties for each building, to determine results for the cohort as a whole.

The R^2 value in this process was set empirically, and the CV(RMSE) threshold was set conservatively, drawing from the ASHRAE Guideline requirements for projects for which uncertainty is *not* computed.

A common question surrounding the use of M&V 2.0 tools is how accurately their embedded models can be used to quantify savings. Modeling error can be quantified with goodness-of-fit metrics such as CV(RMSE). The uncertainty in the savings result that can be attributed to model error is a useful concept in understanding the quality of a whole-building savings analysis. As previously noted, we assess the uncertainty due to model error using the formulation defined in ASHRAE Guideline 14 [13], for models with correlated residuals.

4. Results

4.1 Number of Buildings For Which the Baseline Could Be Determined Automatically

Of the eighty-four commercial buildings for which at least nine months of baseline meter data were available, fifty-four were automatically fit with a baseline model that had an R^2 value above .6, and a CV(RMSE) lower than 25%. All eighty-four buildings were fit with an NMBE that met the ASHRAE guidance (NMBE<0.5%). For the thirty buildings that could not be automatically modeled with an acceptable goodness of fit an inspection of the data was conducted. In seven cases it appeared that the documented measure implementation date was incorrect, i.e., a large reduction in energy consumption was observed during the baseline period, and the load profile was consistent with that observed in the measure post period; in several cases this was confirmed through discussion and review with the utility partner. An example of one of these cases is shown in Figure 2. The majority of the remaining twenty three buildings for which model fitness was poor revealed seasonal trends that were not captured by the baseline model. In these cases it was not possible to identify a straightforward, justifiable adjustment to obtain a well-fit model – some of these buildings may not be well suited for automated M&V2.0, and would require more resource intensive custom investigations by a professional practitioner. Overall, fifty-four of seventy-seven buildings (those with a correct measure implementation date), or 70% of the buildings in the study were found to be well suited to automated characterization of baseline energy use.
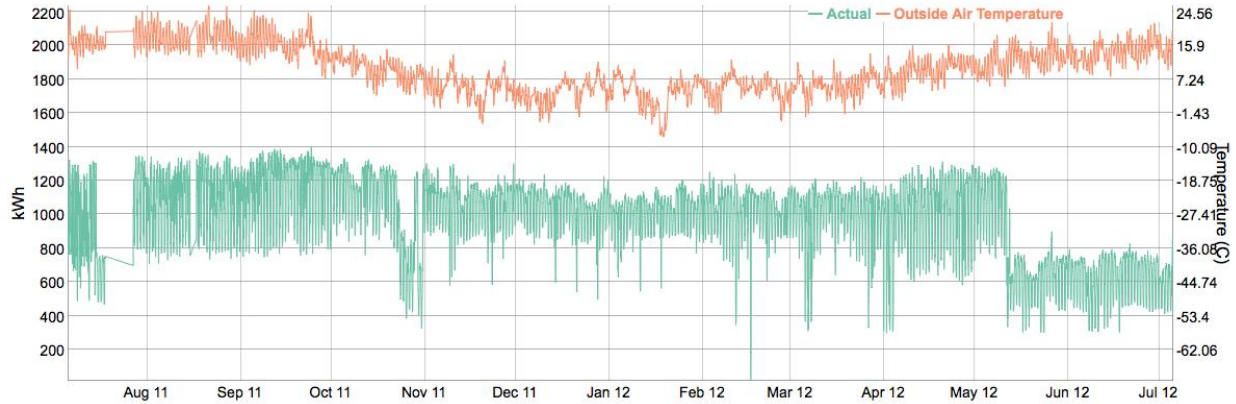
Figure 2. Example of a suspected case of incorrect documentation of measure implementation date; the measure may have been installed in mid-May rather than Mid-July, as was reported.

4.2 Uncertainty and Confidence in Gross Savings, Due to Model Error

For each of the fifty-four buildings that met the R^2, CV(RMSE), and NMBE fitness screening described in *Section 4.1*, and for each set of program data, the uncertainty in savings due to model error was calculated at the 95% confidence level. All fifty-four buildings had at least 9 months of baseline data. For 49 of these buildings 12 months of 'post-measure' data were available, and 5 had at least 11 months of post data.

Figure 3 shows a graphical illustration of the savings and uncertainties results. In this plot each building is represented with a vertical bar. The y-axis shows the fractional, i.e., percent savings that were estimated for each building. The error bars show the fractional savings uncertainty due to model error, also plotted for each building. For a portion of the buildings - those at the left of the plot – the pre/post meter data indicated zero savings, or even increases in consumption after measure implementation. These effects may be due to failed measures, or changes in building loads or operation that are independent of the measures that were implemented.

For Data Set 1 the 10th percentile, the median, and the 90th percentile of FSUs were equal to 9%, 27%, and 180%. For the three buildings in Data Set 2 the FSUs were equal to 18%, 44% and 140%. For Data Set 3 the 10th percentile, the median and the 90th percentile were respectively equal to 14%, 23% and 182%.
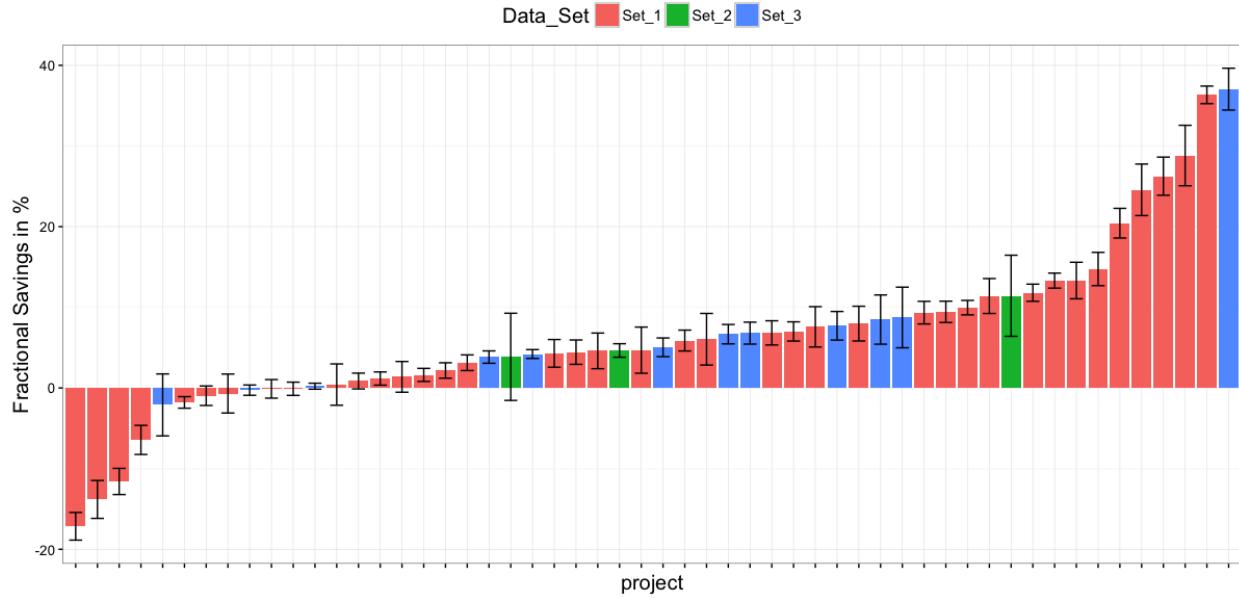
Figure 3. Savings uncertainty ranges for each of 54 buildings, at 95% confidence level

The aggregated uncertainty for each data set was quantified by using uncertainty propagation as described in the *Methodology, Section 3*. Table 2 summarizes these results. In the second column of the table the aggregated fractional savings is provided with the corresponding uncertainty ranges. In the third column the aggregated FSU is indicated for each data set. As a benchmark, the final column in the table lists the fraction of buildings that meet the confidence and uncertainty threshold established in ASHRAE Guideline 14, i.e. 68% confidence, and no more than 50% FSU, which is equivalent to a FSU equal to 100% at 95% confidence level. Overall, aggregating the savings resulted in a decrease in the uncertainty, with the aggregated value falling below that of the best 10th percentile of the individual buildings. In addition, the majority of buildings meet the ASHRAE guidance concerning the precision of the savings estimates.

Table 2. Fractional Savings Uncertainty due to model error the 95% confidence level, and fraction of individual buildings meeting ASHRAE uncertainty guidance

Data Set		Aggregated Fractional Savings with the Uncertainty Range	FSU	Fraction Meeting ASHRAE Guidance	Median of FSU At Building Level
Data Set 1	Screened for model fit, n = 39	[3.66; 3.96; 4.26]	7.6%	82%	27%
Data Set 2	Screened for model fit, n = 3	[4.54; 6.51; 8.47]	30.1%	66%	44%
Data Set 3	Screened for model fit, n = 12	[5.43; 6.14; 6.86]	11.7%	75%	23%

4.3 Comparison of Results from Automated Whole-Building, vs. Traditional M&V

Traditional measurement and verification may be conducted at the measure level with shorter-term system-level measurements or at the whole-building level, with longer-term measurements and custom models that are created for each building. These calculations that are based on measurements after the measure was implemented are referred to as ex-post savings estimates. 'Ex-ante' estimates of *expected* savings are calculated before the measure is implemented. Although a whole-building pre/post savings estimate may not always be directly comparable to, or expected to match an ex-ante or ex-post measure-level estimate, the industry has expressed interest in knowing how the results from automated whole-building tools differ from those based on industry-standard traditional approaches.

To explore how this question might be addressed, we first identified the buildings in the study data sets for which prior savings estimates were available *and* that had passed the baseline fitness test for R^2, CV(RMSE), and NMBE. The data from the post-measure implementation period was visually inspected to determine whether suspected *'non-routine'* events were present. In measurement and verification, non-routine events are defined as changes in building energy consumption that are not due to changes in independent variables in the baseline model (in this study, time of day, day of week and outside air temperature), and that also are not due to the measure itself. If large enough, non-routine events may appear as otherwise anomalous changes in the building's load profile. For accurate savings estimation, these non-routine events must be quantified, and accounted for as '*non-routine adjustments'* in the avoided energy use savings calculation. This is a manual process that requires some degree of engineering expertise, as well as knowledge of what the non-routine event was. For example, commonly encountered non-routine events include changes in building occupancy or schedules. Figure 4 shows an example of a building from the study Data Set 2 in which suspected non-routine events were present (see highlighted area in bottom right of plotted data). Since there was not sufficient information available from the project records to quantify the energy impact of suspected non-routine events, these buildings were not considered for this portion of the analysis.
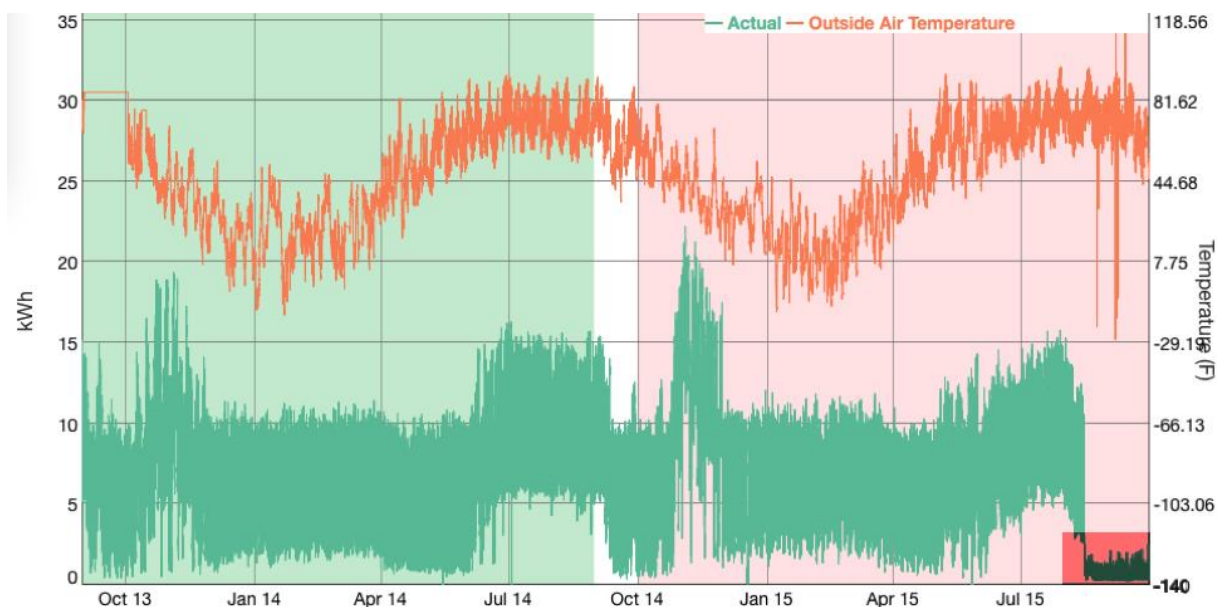
Figure 4. Example of a building with a potential non-routine event after measure implementation

After filtering the full study data set for sufficient time history of pre- and post-measure meter data, baseline model fitness, suspected non-routine events, and the availability of prior savings estimates, there was not a sufficient number of buildings for which to draw conclusions or generalize our findings. However, the results are presented as an illustration of how principles of uncertainty analysis can be used to understand how divergent or similar automated whole-building savings results are from traditional savings estimates. Table 3 summarizes the aggregated ex-post savings estimate for the buildings that remained in each data set, after filtering. It also contains the 95% confidence interval for savings that were estimated with the automated whole-building tool. For data set 3, the ex-ante savings estimated was available, and is also provided for reference.

Table 3. Aggregated ex-ante and ex-post savings estimates

Data Set	Ex-Ante Savings Estimate [kWh]	Ex-post Savings Estimate [kWh]	Automated Whole-Building Savings Estimate, 95% Confidence Interval [kWh]
Data Set 1, n =39	Information not available	Information not available	n/a
Data Set 2, n =3	Information not available	198,316	[225,680; 420,675]
Data Set 3, n= 5	3,041,646	3,038,452	[2,405,841; 3,105,617]

For data sets 3, the ex-post savings that were determined using traditional M&V approaches were within the 95% confidence interval of the savings that were determined using the automated whole-building tool. This indicates that the automated result was statistically equivalent to the result obtained using standard engineering calculations. This was not the case for data set 2, where the ex-post estimate was statistically different from the automated result. To reiterate, given the small size of the data set, one cannot generalize from these findings, however

this type of analysis can be replicated and expanded to address questions related to the comparability of traditional versus automated whole-building savings results.

4.4 Estimation of Labor Time Required for Automated Whole-Building, vs. Traditional M&V

Estimates of the labor time required to conduct the savings analysis in the automated versus traditional M&V case were based on several sources using program Data Sets 1 and 2. For the automated case, the time that it took a researcher to conduct the fitness screening procedure and compute results with the M&V 2.0 tool was tracked and documented. For the traditional case the time requirements were estimated by program implementers. For Data Set 1, based on the measures implemented (retro-commissioning with some retrofit), the traditional approach to M&V would have been a whole-building assessment conducted in a more manual process, without an M&V 2.0 tool. An implementer working with the utility on M&V 2.0 tools estimated that it would take approximately 4 days to quantify the savings with standard tools to review the data, and create custom baseline models for each building. For Data Set 2 (specific measure types not provided for the study), the implementer reported that the traditional approach to M&V that was previously implemented comprised custom engineering calculations. The implementer estimated that it previously took approximately 6 days to quantify the savings.

These findings are summarized in Table 4, and suggest that the use of automated tools has the potential to reduce the labor time that is required to estimate project-level and possibly program-level gross savings. If technical review of the savings estimates are also required, the reduced number of data streams, and use of standardized calculations has the potential to result in further labor time reductions, particularly in the case of engineering calculations. Conversely, the need to monitor for and quantify non-routine adjustment may outweigh these advantages, and these tradeoffs are highlighted as an area of focus for future work. Note that it was not within the scope of this study to estimate the time required to quantify savings for the ~25% of buildings in the data set that were not well-modeled with the automated approach, therefore requiring deeper engineering investigation.

Table 4. Estimates of time required to conduct traditional M&V and automated M&V for the data sets in this study.

Data Set	Time to Conduct Traditional M&V	Description of Traditional M&V Approach	Time to Conduct Automated M&V
Data Set 1	4 days	Manual whole-building M&V	1 day
Data Set 2	6 days	Custom engineering calculations,	1 day
Data Set 3	Information not available	Information not available	Information not available

5. Discussion and Conclusions

The results of this work indicate that automated M&V2.0 methods can be used to accurately quantify whole-building savings, and that automation can offer time savings advantages. The analyses in this study showed that automated tools that use commonly available interval meter and weather data can model baseline energy use at industry standard levels of

fitness for a large fraction of buildings. For the data sets in this work, seventy percent of the buildings were well suited to the automated approach. It was not possible to characterize common traits of the buildings that were not easily modeled, as that information was not available, however prior work [3] suggests that building type alone, is not a conclusive indicator of load predictability. Buildings that are not well fit could be handled in multiple ways. They could be recognized to have a less accurate result, and nonetheless be included in the aggregate, additional explanatory variables could be identified to include in a more custom model, or the savings estimation could be conducted with an alternate method that does not rely on whole-building meter data.

Model fitness in the baseline period, as quantified with metrics such as CV(RMSE) is one gauge of the accuracy of the savings estimation. However a more direct indication of accuracy is the uncertainty in the ultimate savings result that is due to the model error. For the program data sets analyzed in this study, in 80% of the cases, savings and uncertainties for each individual building were accurate to levels above the ASHRAE Guideline. When the buildings were pooled and treated as a portfolio, the accuracy of the result improved, i.e., the uncertainties were reduced significantly. It is important to note that these uncertainties are those attributable solely to errors in fitting the model to the baseline data – M&V protocols and guidelines do not address the uncertainty associated with non-routine adjustments that may be made to account for effects such as changes in occupancy or building scheduling. These uncertainties are likely larger than those associated with baseline model error.

While uncertainty is not commonly considered today, it could hold value for evaluating and reducing project and investment risk. For example, ASHRAE's published methods for computing fractional savings uncertainty depend on depth of savings, length of the training and prediction periods, and model CV(RMSE). "Look-up" tables can be used to explore the likelihood that a given model will produce savings estimations that meet uncertainty and confidence requirements, for a specific set of buildings and expected depth of saving. Uncertainty-based approaches are also valuable because they permit an understanding of the risk of under-performance. They align nicely with approaches used by the financial industry, in which investors are accustomed to decision-making under uncertainty and increased use of these concepts could potentially facilitate more effective project financing by private investors.

Since uncertainty analysis has not commonly been used in the industry to report gross savings, only point value savings estimates were available for the historic data sets analyzed. Therefore a direct comparison of accuracy between the automated and traditional approaches was not possible. However to illustrate how one might analyze how 'different' the results are, we compared the aggregated ex-ante and ex-post savings from the traditional method to the range of savings from the automated savings, at the 95% confidence level. Although the data sets were too small to generalize from, analyses of this type will be important in determining whether results are indeed statistically different from one another, and in turn, whether the use of automated savings analyses is likely to significantly change the industry's current accounting frameworks.

While it is possible to calculate savings uncertainty due to model error, the industry has not yet articulated clear acceptance criteria to define the bar for rigor that must be met. For example, what are acceptable levels of confidence and uncertainty? How must results and the process be documented, and what elements must be transparent? With agreed upon acceptance criteria, regulators could require the use of M&V methods that met these criteria, setting a clear

bar for evaluation. This could inform further conversation about the demonstrated accuracy and uncertainty of known methods for a given program type. By agreeing to these criteria and methods up front, evaluation risk and ratepayer risk could be diminished. An approach to energy efficiency that includes M&V methods with known uncertainty may also facilitate the effective integration of energy efficiency into state efforts to measurably reduce greenhouse gas emissions by placing efficiency-based GHG reductions in a robust reporting framework similar to that of supply-side resources.

Future Work

Future work will focus on four key areas necessary to support the broader adoption of these streamlined approaches to M&V. First, facilitation and analysis of in-situ pilots in 'live' projects. The current work that analyzed historic data, was an important first step, however was limited due to the fact that the data was not originally acquired with an objective of assessing automated meter-based approaches to M&V. Pilots will enable more thorough testing of the feasibility of these techniques, focusing on the benefits of continuous feedback, practitioner workflows, and side-by-side comparisons of accuracy, time and cost relative to traditional approaches. These pilots will be conducted in partnership with utility implementers, evaluators, and the regulatory community, and should include both commercial and residential applications.

A key outstanding issue related to continuous feedback, accuracy, and feasibility, concerns the treatment of non-routine adjustments. Appropriate treatment of these changes in energy use that are not due to the measure, and not captured in the model, is critical to achieving an accurate savings result, and the ease of doing so has important implications for labor time savings potential. Today these adjustments are identified manually, however, if their detection can be automated, then cost and accuracy can be further improved. New ways of measuring high-impact independent variables will also reduce the complexity of handling non-routine events. For example, if occupancy sensing could be scaled, and made as accessible as weather data for use an independent variable, the performance of automated tools could be significantly improved, and one of the more common sources of non-routine events (changes in occupancy) could be eliminated.

The third area of future work will entail continued engagement with the evaluation and regulatory communities to establish acceptance criteria for M&V 2.0. This will include uncertainty and confidence requirements, as well as documentation needs for transparent third party review. For example, while the analyses in this work were conducted at the 95% confidence level, the industry has yet to define confidence and associated uncertainty levels that must be for results to be considered sufficiently rigorous. Similarly, qualitative requirements for documentation must be set. Closely related, there is a need to explore improved approaches to uncertainty quantification that may better estimate the confidence interval, and that can be applied to a wider range of models such as machine leaning and highly non-linear techniques.

Finally, today's M&V 2.0 tools focus on existing conditions baselines, and do not quantify savings versus more commonly to-code or industry standard practice baselines. As these software-based approaches are further tested, and their viability proven for the existing conditions case, there may be cause to extend their capabilities, expanding their applicability to a wider variety of program designs and measure types.

Acknowledgement

References

[1] Granderson, J, Piette MA, Ghatikar, G. 2011. Building energy information systems: User case studies. Energy Efficiency 4(1): 17-30.

[2] Granderson, J, Lin, G. 2016. Building energy information systems: Synthesis of costs, savings, and best-practice uses. Energy Efficiency 9(6), 1369-1384.

[3] Granderson, J, Price PN, Jump, D, Addy N, Sohn, MD. 2015. Automated measurement and verification: Performance of public domain, whole-building electric baseline models. Applied Energy 144: 106-113.

[4] Granderson, J, Touzani, S, Custodio, C, Sohn, M.D, Jump, D. and Fernandes, S, 2016. Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings. Applied Energy 173: 296-308.

[5] Kramer, H, Effinger, J, Crowe E. Energy management and information system (EMIS) software technology assessment: Considerations for evaluating baselining and savings estimation functionality. Pacific Gas and Electric, 2013, ET Project Number ET12PGE1311.

[6] Kramer, H, Russell, J, Crowe, E, Effinger, J. Inventory of Commercial Energy Management and Information Systems (EMIS) for M&V Applications, Northwest Energy Efficiency Alliance, 2013, #E13-264.

[7] Reddy, TA, Saman, NF, Claridge, DE, Haberl, JS, Turner WD, Chalifoux AT. 1997. Baselining methodology for facility-level monthly energy use – Part 1: Theoretical aspects. AHSRAE Transactions 103(2):336-347.

[8] DNV-GL. The changing EM&V paradigm: A review of key trends and new industry developments, and their implications on current and future EM&V practices. Report prepared by DNV-GL for the NEEP Regional EM&V Forum, December 2015.

[9] Rogers, E, Carley, E, Deo, S, Gossberg, F. How information and communication technologies will change the evaluation, measurement, and verification of energy efficiency programs. American Council for an Energy Efficient Economy, 2015, #IE-1503.

[10] Kupser, J, Francois, S, Rego, J, Steele-Mosey, R, Galvin, T and McDonald, C, 2016. M&V 2.0: Hype vs. Reality. Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings. August 2016.

[11] Efficiency Valuation Organization (EVO). International Performance Measurement and Verification Protocol: Concepts and options for determining energy and water savings, Volume I. January 2012. EVO 10000-1:2012.

[12] Granderson, J, Price, PN. 2014. Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models. Energy, 66, 981-990.

[13] ASHRAE Guideline 14 (2014). ASHRAE Guideline 14-2014 for Measurement of Energy and Demand Savings, American Society of Heating, Refrigeration and Air Conditioning Engineers, Atlanta, GA.

Appendix

The R^2 corresponds to the proportion of the energy use variance explained by the model. The R^2 value ranges between 0 and 1, with 0 indicating that the model explains none of the output variability, and 1 indicating that the model explains all the output variability. It is defined as:

$$R^2 = 1 - \frac{\frac{1}{n}\sum_i^n(y_i - \hat{y}_i)^2}{var(y)}$$

In this equation y_i is the actual metered value, $\hat{y}_i$ is the predicted value, n is the total number of data points and *var(y)* the variance of the actual metered value.

The NMBE quantifies the total difference between model predicted energy use, and actual metered energy use. The NMBE is defined in the following equation, where $\bar{y}$ is the average of the y_i:

$$NMBE = \frac{\frac{1}{n}\sum_i^n(y_i - \hat{y}_i)}{\bar{y}} \times 100$$

According to this convention, if the value of NMBE is positive, the prediction of the total energy used during the prediction period is lower than the measured value. A negative NMBE indicates that that the prediction is higher. The value of NMBE is independent of the timescale on which it is evaluated, which means that the value of the metric will be the same if the timescale is 15-minute, hourly or daily.

The CV(RMSE) provides a quantification of the typical size of the error relative to the mean of the observations. This metric also indicates the model's ability to predict the overall load shape that is reflected in the data. CV(RMSE) is also familiar to practitioners, and is prominent in resources such as ASHRAE Guideline 14. The CV(RMSE) is defined below, where y_i is the actual metered value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the average of the y_i, and n is the total number of data points:

$$CV(RMSE) = \frac{\sqrt{\frac{1}{n}\sum_i^n(y_i - \hat{y}_i)^2}}{\bar{y}} \times 100$$

In contrast to the NMBE, R^2 and CV(RMSE) quantify the predictive accuracy at the timescale of the data and prediction.